Robust Machine Learning Algorithmic Rules for Detecting Air Pollution in the Lower Parts of the Atmosphere

RESEARCH PAPER

KASSIM MWITONDI
HUGO WAI LEUNG MAK
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Sophisticated data-intensive approaches have been widely applied in addressing air pollution problems, with applications ranging from remote sensing quantification of ground-level concentrations of atmospheric pollutants to associating particulate matter with atmospheric CO_2. The biggest challenge to such applications, however, remains model optimisation—a problem that derives from inherent randomness in training, validation and test data. A standard approach to address data randomness hinges on data harmonisation and data augmentation—two concepts that naturally appeal to the highly "non-orthogonal" 17 Sustainable Development Goals (SDG). This paper proposes a novel approach with built-in robust mechanisms for generating the "most parsimonious model" – with potential "global representativeness" to highlight data-driven solutions of regional and global environmental challenges. The proposed approach is powered by two algorithms that sequentially estimate, maximise and optimise parameters from thirty thousand ground-level air pollution data points obtained from different locations in southern China; generate statistical associations among the pollutants and present interpretable visual outputs. The algorithms balance the power of data, machine learning techniques and underlying domain knowledge to enhance problem identification and solution development. The results show optimal associations between spatio-temporal attributes and relevant pollutants, thus provide useful insights into the state of pollution in southern China. The findings also indicate robustness of features that exhibit a great potential for building analytical bridges across disciplines and sectors. This research is expected to contribute to our understanding of how pollutants are spatially distributed within the lower part of the atmosphere, potentially leading to improved model performance and innovation. Further, it will also contribute to the design of methods to deal with challenges posed by the "non-orthogonality" of socio-economic, technical and environmental attributes of the SDG.

CORRESPONDING AUTHOR:
Hugo Wai Leung Mak
Dept of Mathematics, The Chinese University of Hong Kong, Hong Kong, China; Dept of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China
mahwlmak@ust.hk

KEYWORDS:
Air Pollution; Association Rules; Correspondence Analysis; EM Algorithm; K-Means; Machine Learning; Principal Component Analysis (PCA); Robust Estimation; Sustainable Development Goals (SDG)

TO CITE THIS ARTICLE:
Mwitondi, K. and Mak, H.W.L. (2025) Robust Machine Learning Algorithmic Rules for Detecting Air Pollution in the Lower Parts of the Atmosphere. *Data Science Journal* 24: 27, pp. 1–24. DOI: https://doi.org/10.5334/dsj-2025-027

1 INTRODUCTION

Data-intensive approaches, pertaining to machine learning and artificial intelligence, have been widely applied in addressing air pollution problems. Typical examples include data clustering and association rule mining methods, which have found applications in different settings (Soares *et al.*, 2023; Wu, Wen and Zhu, 2024). The former, for instance, have been applied in quantifying pollution levels, particularly in remote sensing, where they have been applied to obtain ground-level concentrations of atmospheric pollutants. Examples of the application of the latter include the association of particulate matter with atmospheric CO_2 levels using the measurement of one pollutant to estimate the concentration of another. However, most approaches are still hampered by model optimisation challenges that derive from inherent randomness in spatio-temporal and sectoral data variations. A standard approach to address data randomness hinges on the concepts of data harmonisation (Cheng *et al.*, 2024; Diaz-de Arcaya *et al.*, 2025) and data augmentation (Mumuni, Mumuni and Gerrar, 2024; Wang *et al.*, 2024). The former describes a process for resolving variations in data syntax, while the latter relates to artificially enhancing data size, diversity and other parameters in order to improve model performance. This paper proposes a novel approach to address data randomness, with built-in robust mechanisms to generate the "most parsimonious model" with a potential "global representativeness" to highlight data-driven solutions to regional and global environmental challenges. It is powered by two algorithms that sequentially estimate, maximise and optimise parameters from thirty thousand ground-level air pollution data points obtained from different locations in southern China; it generates statistical associations among the pollutants and presents interpretable visual outputs.

1.1 PURPOSE OF RESEARCH

The main purpose of this research is to develop a data modelling method to address societal challenges, such as air pollution – a local, regional and global multi-dimensional challenge, that cuts across disciplines, sectors and regions. Despite being one of the greatest environmental threats to human health and well–being, air quality is not explicitly highlighted in the United Nation's Sustainable Development Goals (SDG) agenda (Zusman, Elder and Sussman, 2020). It appears subtly in several SDGs, for instance, the interaction between climate change and air pollution has been reported to "amplify risks to human health and crop production" (Sillmann *et al.*, 2021), and air pollution has been linked to respiratory and cardiovascular diseases (Mak and Ng, 2021; Newby *et al.*, 2015; Zhang *et al.*, 2022b). Ozone, for instance, is known to reduce crop yields, implying that addressing its spread supports food security. The above examples are related to the SDGs 1, 2, 3, 8, 9, 10 and 11. More specifically, target 11.6.2 of SDG 11 – sustainable cities and communities – focuses on reducing the environmental impact of cities by improving surrounding air quality conditions. We are therefore challenged to adopt interdisciplinary approaches to uncover the triggers of SDG indicators, by detecting information hidden in their data attributes (Mwitondi and Said, 2021; Mwitondi, Munyakazi and Gatsheni, 2018b).

Thus, the choice of air pollution data for developing and testing the method lies on the fact that "it affects, and/or is affected by, almost all the 17 SDGs". Most importantly, the SDG epitomise all aspects of our existence and so, it is reasonable to assume that through coordinated and collaborative studies on air quality, the research community can unfold our understanding of how they interact and the impact of such interaction on our livelihood. In other words, it will help unify spatio-temporal and other global initiatives, such as the World Health Organization (WHO) air quality guidelines (WHO, 2021), United Nations Environment Programme (UNEP) reports (UNEP, 2023), and other organisations such as the Global Alliance on Health and Pollution (GAHP, 2023).The sampled dataset is assumed to reflect "global representativeness" in air pollution modelling, given China's role on both sides of the pollution spectrum as well as the large number of environmental studies that have been carried out in the country.

1.2 STUDY MOTIVATION

The motivation for this work derives from our research interest in addressing societal challenges through Big Data Modelling of Sustainable Development Goals (BDMSDG) (Mwitondi and Said, 2021; Mwitondi, Munyakazi and Gatsheni, 2020). The interactions of humans and their habitat inevitably lead to mutual impacts that often turn out to be adverse on both sides (Anser *et al.*,

2024; Edo *et al.*, 2024). The paper is set on the premise that human and natural activities generate large volumes of multifaceted environmental and other types of data, much faster than the rate at which the data can be processed, hence the need for developing more sophisticated, faster and more efficient data analytic tools and methods than ever before. It focuses on the applications of machine learning techniques in environmental modelling, which is a fundamental aspect of SDG. Thus, It is reasonable to assert that the motivation hinges on the complexity of SDG and the way to address them through robust data modelling methods.

1.2.1 Complexity of SDG Interactions

For an intuition into the mutual impact of human–nature interactions, consider the inherent correlations among different aspects of the 17 SDGs (United Nations, 2015). One aspect that relates to each SDG is "environmental pollution". Thus, conducting air quality studies and spatial analyses can play an important role in planning and laying down policies for land use, infrastructural and business development, which hinge on different SDGs. For instance, if pollution hotspots in southern China (a major trading port and manufacturing hub) can be identified, preventive measures and stricter pollution abatement policies can be implemented in concerned cities or counties, so that overall pollution levels can be effectively reduced, ensuring healthy lives and promoting well–being for all citizens and visitors (SDG #3).

Air quality and climatic conditions mutually impact each other: climate change can affect air quality, and certain air pollutants can affect climate change (Feng *et al.*, 2019) (SDG #13). Other SDGs that are closely associated with air quality include SDG #11 (Sustainable Cities and Communities), the attainment of which relates to reducing air pollution and making cities and human settlements inclusive, safe, resilient and sustainable. Smart and sustainable city development shows care and a focus on maintaining good air quality within neighbourhood levels and within various spatial scales. Focusing on a specific region, such as southern China, contributes to SDG #17 (Partnerships), which facilitates interdisciplinarity across geographical regions, legislations, infrastructural setups, and promotes citizen engagement.

1.2.2 Addressing the Complexity of SDG interactions

Recognising the complexity stated in Section 1.2.1, we adopt the mathematical concept of "non-orthogonality" of SDGs (Mwitondi, Munyakazi and Gatsheni, 2020). Different methods have been developed to track, monitor and model environmental pollution, including automated data-driven tools to capture, model and track environmental variations (Mwitondi, Munyakazi and Gatsheni, 2020). The literature on environmental pollution is awash with applications of sophisticated statistical and machine learning methods (Liu *et al.*, 2022a; Pan, Harrou and Sun, 2023). Data–intensive, machine learning approaches have been used in the remote processing of ground-level concentrations of atmospheric pollutants such as NO_2, $PM_{2.5}$ and ozones (Chi *et al.*, 2022; Du *et al.*, 2022; Xu *et al.*, 2018). Clustering methods have been applied in quantifying pollution levels as well as in remote sensing (Hsu *et al.*, 2023; Zhang and Yang, 2022). Other studies have applied temperature inversion methods (Feng, Wei and Wang, 2020), partial differential equations with appropriate initial and boundary conditions (Shafiev, 2024), as well as the use of remotely sensed datasets (Lin *et al.*, 2020; Mak *et al.*, 2018). A study focusing on surface and sea surface pressure, geopotential height, temperature, relative humidity, wind field, and vertical velocity, established an association between regional weather and climate events on one side and large-scale circulation anomalies on the other (Cai *et al.*, 2020). Global studies on the relationship between economic development and pollution are also well documented (Yan *et al.*, 2024). The main challenge of the foregoing studies revolves around model optimisation (Liu *et al.*, 2022a; Vardoulakis *et al.*, 2007; Zhang *et al.*, 2022a), i.e., the need to develop "robust" methods to perform relevant tasks that can be replicated in a spatio-temporal context. Further, socio-demographic factors, coupled with economic, cultural, political, legislative and technical variations, directly affect pollution levels as well as mitigation efforts. Focusing on southern China was therefore particularly appealing due to the country's role on both sides of the pollution spectrum, i.e., as a polluter and as a pollution mitigator (Liu *et al.*, 2022b; Sun, 2016; Zheng and Kahn, 2017). Furthermore, most of the aforementioned studies have been carried out in China, hence associating results from a "robust" model with "acceptable metrics" such as SDG indicators and sharing the relevant datasets publicly would be a major step towards our understanding of SDG attainment (Mak and Lam, 2021; Mwitondi, Munyakazi and Gatsheni, 2020).

Our paper complements previous studies by adopting an interdisciplinary approach to spatio-temporal air pollution modelling, via the two algorithms in Section 2.2.4 that are embedded with the flexibility to run different techniques including, in this specific application, Principal Component Analysis (PCA), Cluster Analysis (Chapmann, 2017; Kogan, 2007), Correspondence Analysis (CA) (Hirschfeld, 1935), K-Means (Lloyd, 1957; MacQueen, 1967) and the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977). Our main idea hinges on an interdisciplinary approach to both problem identification and solving, i.e., balancing the power of data, machine learning techniques, and underlying domain knowledge. We focus on the identification of pollution levels of different pollutants in the sampled area, then propose two algorithms for identifying spatio-temporal associations among key air pollutants. The novelty of this study derives from the role of the two algorithms, which hinges on their "robustness" in addressing data randomness and on highlighting the path towards addressing the "non-orthogonality" of SDGs, i.e., the factors that affect the extent and impact of air pollution within any geographical area are not confined to the geographical boundaries of that area. This can be better understood in the context of how our atmosphere is structured, because pollutants along the air can travel from one country to another within a short period of time, as well as across various vertical layers within the atmosphere. Table 1 describes the structured layers of our atmosphere and their relevance to humanity.

Table 1 Layers of our atmosphere.

LAYER	DESCRIPTION & RELEVANCE TO HUMANITY
Troposphere	Closest to our habitat–stretching up to 10 km above earth. Its temperature decreases inversely with distance from the centre of the earth (approx. 6.5°C per kilometre) (Omrani *et al.*, 2022).
Stratosphere	Consists the majority of atmospheric ozone, which absorbs ultraviolet radiation and protects us from potential health risks. It is characterised by high temperatures over summer and lowest over the winter period (Xu *et al.*, 2023).
Mesosphere	The temperature varies inversely with vertical height above ground (Laštovička, 2023).
Thermosphere/ Ionosphere	Absorption of energetic ultraviolet & X-ray radiation from the sun, thus temperature increases with vertical height. They also vary between night and day as well as between seasons. It reflects and absorbs radio waves, allowing global radio wave transmission (Goncharenko *et al.*, 2021).
Exosphere	Contains mainly oxygen and hydrogen atoms, but they rarely collide - they follow "ballistic" trajectories under the influence of gravity (Janches *et al.*, 2021).
Magnetosphere	The outer region surrounding the earth, where charged particles spiral along the magnetic field lines, with the earth behaving like a huge magnet (Lu *et al.*, 2022).

In this study, we focus on the tropospheric layer – the bottom layer of the earth's atmosphere. It constitutes of about 75–80% of the atmospheric mass, with its temperature variation affected by height and time of day and/or year. Most of our terrestrial weather—clouds, rain, snow—derive from it, making it a particularly interesting research scope.

1.3 RESEARCH QUESTION AND OBJECTIVES

Identification of relevant data attributes and the nature of their complex interactions are fundamental to creating robust data-driven solutions. Using air pollution data, described in Table 2, this study seeks to address the following problem: ***Identifying optimal parameters for air pollution control by learning rules from datasets of multiple pollutants.*** Learning rules from data is particularly important if our data processing capabilities are to match the rate at which we generate it (Ridzuan and Zainon, 2022; Zhou *et al.*, 2021). To address the problem, we lay down the following general and specific objectives.

1. **To develop a robust data-driven method for air quality monitoring and control**
   - **(a)** To harmonise data from disparate air quality stations in southern China (including Hong Kong and Macau).
   - **(b)** To clean and prepare the newly created composite dataset for large-scale modelling.
   - **(c)** To design and test mechanisms for extracting and analysing information relevant to the research problem.

2. **To apply the established method on datasets acquired from multiple stations in southern China (including Hong Kong and Macau)**
   - **(a)** To carry out initial exploratory data analysis to understand the general behaviour of the data.
   - **(b)** To identify associations among data attributes like pollutants, timelines and locations.
   - **(c)** To select and optimise key parameters and match patterns in a spatio-temporal context.
   - **(d)** To demonstrate and assess applicability of the method across other SDG applications.

2 METHODS

This section introduces data sources and techniques used in the study. The adopted methods are twofold, namely, technical and applied. From the technical perspective, we present two algorithms. One "estimates" the parameters of the pollutants and the other uses the estimates to perform comparative analysis for best model selection. The two algorithms are applied to establish optimal associations among selected air pollutants, as well as between pollutant types at discretised daily and annual time periods. From the application perspective, it hinges on improving the quality of life by addressing air pollutant trends, hence touching on several SDGs and their "non-orthogonality".

2.1 DATA SOURCES

Pollution datasets were collected from southern China, including Hong Kong and Macau. Data came from 272 stations that included ambient, roadside, and environmental resources monitoring units, with common collection sites including schools, monitoring sites, parks, roadside, street entrances, etc. The data covered the period 00:00hrs on 1 January 2023 to 23:00hrs on 31 January 2023, 00:00hrs on 1 April 2023 to 23:00hrs on 30 April 2023, 00:00hrs on 1 July 2023 to 23:00hrs on 31 July 2023, and 00:00hrs on 1 October 2023 to 23:00hrs on 31 October 2023, with a spatial coverage of longitude from 109°E to 117.5°E and latitude from 19.5°N to 25.5°N, as illustrated in Figure 1. The spatio-temporal coverage was intended to capture the seasonal and spatial variations in prescribed time periods and geographical contexts.

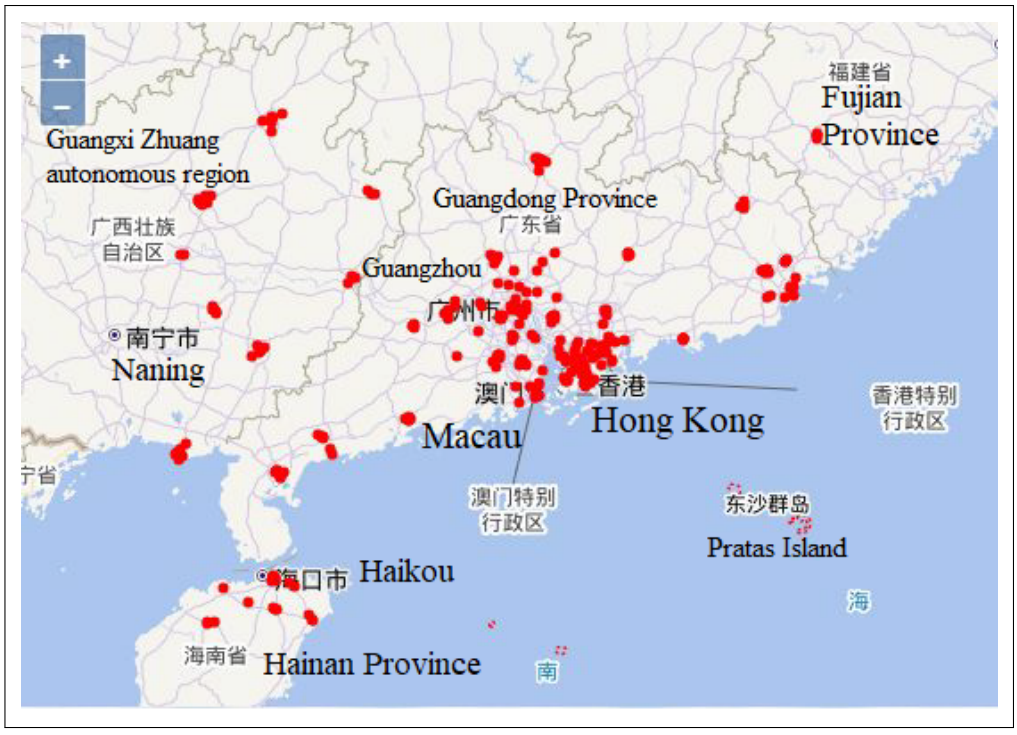

Figure 1 Data sources in southern China including Hong Kong and Macau.

More specifically, the dataset was collected from the Air Quality–China National Environmental Monitoring Center (AQ–CNEMC), the Hong Kong International Airport (AQ–HKIAWEB), the Air Quality Management Information System (AQMIS) and the Macao Meteorological and

Geophysical Bureau (AQ–MASMG). The dataset, detailed in Table 2 consists of five pollutants: Respirable Suspended Particulates (RSP), Fine Suspended Particulates (FSP), Nitrogen Dioxide (NO_2), Ozone (O_3) and Nitrogen Oxides (NO_X). The Nitric Oxide (NO) and Nitrogen Oxides (NO_X) ≈ NO + NO_2) data was collected from 15 general and 3 roadside stations set up by the Environmental Protection Department (EPD) in different districts of Hong Kong, providing hourly, daily, monthly and annual average concentrations of each pollutant.

Table 2 Selected key air pollutants.

DATA ITEM	DESCRIPTION	DIMENSION & COMPLETENESS
PM_{10} (FSPMC)	Fine Suspended Particulates (FSP)	2952 × 27 samples: 3.17% missing
NO_2	Nitrogen Dioxide	2952 × 28 samples: 2.18% missing
NO_x = NO + NO_2	Nitrogen Oxides–in Hong Kong	2952 × 19 samples: 2.67% missing
O_3	Ozone	2952 × 114 samples: 4.39% missing
$PM_{2.5}$ (RSPMC)	Respirable Suspended Particulates (RSP)	2952 × 27 samples: 7.9% missing
Time	Daily and hourly recording	From 00:00hrs to 23:00hrs of 1–31 January 2023, 1–30 April 2023, 1–31 July 2023 and 1–31 October 2023 (inclusive)
DayTimes	Discretised time periods of day	• Night: 00:00-06:00hrs • Morning: 06:00-11:00hrs • Day: 12:00-18:00hrs • Evening 19:00-23:00hrs
Period	Monthly weather periods	January, April, July and October

Each of the data items PM_{10}, NO_2, NO_X, O_3 and $PM_{2.5}$ was sampled from multiple stations. The period discretisation into January, April, July and October, was given by design, reflecting the seasonal periods of winter, spring, summer and autumn. Day discretisation into night, morning, day and evening formed part of our data preparation for modelling. The proportions of missing data for each pollutant were negligible, thus relevant observations could be deleted without affecting the probability or spatial distributions, i.e., the distributions were consistent before and after deletion.

2.2 MODELLING STRATEGY

Our modeling strategy is based on the objectives described in Section 1.3. Exploratory data analysis of the harmonised data provides useful insights into the choice, design and development of relevant modelling approaches to address the research problem. Identifying associations among data attributes like pollutants, timelines and locations, as well as estimating and optimising key air quality parameters across samples highlight the paths towards parameter optimisation and replicability of the chosen methods across the SDG spectrum, which addresses part of our study objectives. Thus, the strategy derives from known theoretical aspects of maximisation and optimisation, which are covered in Sections 2.2.1 through 2.2.3. However, given the random nature of training, validation and testing data, the obvious option is to seek "robust estimates" of relevant parameters, which is what the two algorithms in Section 2.2.4 are designed to deliver.

2.2.1 Computational Maximisation of the Log-Likelihood

Some aspects of data visualisation are presented in Section 3.1 to convert a complex scenario into an easy one, which is useful for readers with limited knowledge of statistics. The popularity of data visualisation across data science applications is well documented (Liang *et al.*, 2023; Pika *et al.*, 2021). However, data visualisation has its limitations, particularly when the interest lies on representing characteristics of high-dimensional data. Under such circumstances, density estimation methods, maximum likelihood and posterior estimations are the most appropriate methods.

It is reasonable to assume that the collated data variables are independent and identically distributed, i.e., they are mutually independent, with the same probability distribution. Let us

denote the five pollutants by K simple distributions

$$\phi(x) = \sum_{k=1}^{K} \pi_k \phi(x);\ 0 \le \pi_k \le 1;\ \sum_{k=1}^{K} \pi_k = 1 \tag{1}$$

where ϕ_k represent basic distributions and π_k are group proportions or mixture weights. This implies that we can describe multimodal distributions of any dimension, which we would otherwise only clearly visualise in a 1, 2 or 3-dimensional space. Consider the components of a Gaussian Mixture Model, each with parameters μ_k and Σ_k.

$$\phi(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k);\ 0 \le \pi_k \le 1;\ \sum_{1}^{K} \pi_k = 1 \tag{2}$$

where $\theta = \{\pi_k, \mu_k, \Sigma_k;\ k = 1, 2, ..., K-1, K\}$ are the free parameters of the distribution. Let us denote the pollution data in Table 2 by $\mathbf{X} = \{x_1, x_2, ...x_N\}$ where $x_n,\ n = 1, 2, ..., N-1, N$ are independent and identically distributed from some unknown distribution $\phi(x)$. It can be shown that Equation 2 leads to a set of dependent simultaneous equations that can be solved iteratively. To estimate the aggregate distribution representing density distributions such as those exhibited in Figure 5, we can initialise K mixture components as follows:

$$\phi_k(x) = \mathcal{N}(x|\bar{x}_k, \sigma_k),\ k = 1, 2, ..., K \tag{3}$$

Given $\bar{x}_k$ and σ_k, the estimated aggregating density in Equation 2 will depend on the weights (π_k) assigned to each of the K components, as π_k defines prior memberships to each–i.e., prior probability of the type of pollution. The maximum likelihood estimate of the free parameters θ_{ML} is given as follows:

$$\phi_k(\mathbf{X}|\theta) = \prod_{n=1}^{N} \phi(x|\theta),\ \phi(x_n|\theta) = \sum_{n=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{4}$$

where each of the likelihood terms is assumed to be a Gaussian Mixture Model density, with the log likelihood

$$\log \phi_k(\mathbf{X}|\theta) = \sum_{n=1}^{N} \log \phi_k(x|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) = \mathcal{L} \tag{5}$$

For a single Gaussian model, the sum over K vanishes, and Equation 5 reduces to

$$\log N(x|\mu, \Sigma) = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log \det(\Sigma) - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \tag{6}$$

The parameters that maximise the log likelihood in Equation 5 are obtained by maximising the derivatives of the mean (μ), variation (Σ) and group proportions (π) with respect to θ, as follows

$$\frac{\partial \mathcal{L}}{\partial \mu_k / \partial \Sigma_k / \pi_k} = 0^T/0/0 \longleftrightarrow \sum_{n=1}^{N} \frac{\partial \log p(x_n|\theta)}{\partial \mu_k / \partial \Sigma_k / \pi_k} = 0^T/0/0 \tag{7}$$

Note that the expression in Equation 7, i.e., the necessary conditions for maximising Equation 5, are of the form

$$\frac{\partial \log p(x_n|\theta)}{\partial \theta} = \frac{1}{p(x_n|\theta)} \frac{\partial p(x_n|\theta)}{\partial \theta} \tag{8}$$

where $\theta = \{\mu_k, \Sigma_k, \pi_k;\ k = 1, 2, ..., K\}$ are model parameters, $p(x_n|\theta)$ is the probability of data x_n, given θ, hence

$$\frac{1}{p(x_n|\theta)} = \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \tag{9}$$

The expressions in the denominator in Equation 9 imply that the probability of data given parameters is proportional to the sum of all the K components. It can be further shown that

$$\frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \propto p(x_n|\mu_k, \mu_k, \Sigma_k) = \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{10}$$

The left-hand side of Equation 10 is a normalised probability vector that represents the probability that the data (x_n) was generated by the k^{th} mixture component, which is directly

proportional to the likelihood on the right-hand side. Direct maximisation of the likelihood function of data in the form of a random sample is challenging, so the likelihood and parameters in Equations 1 to 10 can only be obtained by iteratively searching for parameters that maximise $\mathcal{L}$.

2.2.2 Maximum-Likelihood Parameter Estimation and Maximisation Algorithm

If we envision each observation as being characterised by a parametric finite mixture density, we can treat group membership as missing data and use an adapted version of the EM algorithm to estimate and maximise the parameters μ_k and Σ_k. Let $c_i = 1, 2, \ldots, K$ be initial classes with an unobservable indicator variable

$$z_{ik} = \{0, 1\}^K = (z_{i1}, z_{i2}, z_{i3} \ldots z_{iK}) \text{ such that } \begin{cases} 1 & \text{if } c_i = k \\ 0 & \text{if } c_i \neq k \end{cases} \tag{11}$$

We can then compute class membership as the probability of class given data, as follows

$$p(c_i = K|x_i) = p(z_{iK} = 1|x_i) = \frac{\pi_k \phi_k(x_i)}{\sum_{k=1}^{K} \pi_k \phi_k(x_i)} \tag{12}$$

If class membership, central tendency, and variation parameters were observable, they could be estimated as

$$\hat{\pi}_k = \frac{\sum_1^N z_{ik}}{N};\ \hat{\mu}_k = \frac{\sum_1^N z_{ik} x_i}{\sum_1^N z_{ik}};\ \hat{\sigma}_k^2 = \frac{\sum_1^N z_{ik}(x_i - \hat{\mu}_k)^2}{\sum_{k=1}^N z_{ik}} \tag{13}$$

where $\hat{\pi}_k$ is the estimated proportion of data points in class k, $\hat{\mu}_k$ is the mean within that class and $\hat{\sigma}_k^2$ is the variation in that category. However, the parameters are not observable, but we can estimate them from data, at each E step

$$\hat{\pi}_k^{(m+1)} = E\left[z_{ik} = 1|x_i,\ \hat{\mu}_k^{(m)},\ \hat{\pi}_k^{(m)}\right] = \frac{f_k\left[x_i, \hat{\mu}_k^{(m)}, \hat{\sigma}_k^{2(m)},\ \hat{\pi}_k^{(m)}\right]}{\sum_{k=1}^K x_i, \hat{\mu}_k^{(m)}, \hat{\sigma}_k^{(2m)},\ \hat{\pi}_k^{(m)}} \tag{14}$$

The estimated parameters are then maximised at the M step as follows

$$\hat{\mu}_k^{(m+1)} = \frac{\sum_1^N \hat{\pi}_{ik}^{(m+1)x_i}}{\sum_{n=1}^N \hat{\pi}_{ik}^{(m+1)}},\ \hat{\sigma}_k^{2(m)} = \frac{\sum_{n=1}^N \hat{\pi}_{ik}^{(m+1)}\left(X_i - \hat{\mu}_k^{(m+1)^2}\right)}{\sum_{n=1}^N \hat{\pi}_{ik}^{(m+1)}} \tag{15}$$

2.2.3 Extracting Naturally Arising Groups and Components

The estimated and maximised parameters can be used to guide both unsupervised and supervised modelling and in both cases, they potentially help to avoid over-fitting. We adopt PCA and data clustering for dimensionality reduction. The former seeks to transform a number of correlated variables into a smaller number of uncorrelated variables, called principal components, i.e., we can explain the variance-covariance structure of a high dimensional random vector through a few linear combinations of the original component variables. The variables in Table 2 can be formulated in a generic form as in Equation 16, where each extracted component is estimated as a weighted sum of the variables

$$\mathcal{PC}_k = \{w_{ik}FSPMC, w_{ik}NO_2, w_{ik}NO_X, w_{ik}O_3, w_{ik}RSPMC\} \in \mathbb{R}^n \tag{16}$$

Here, $k = 1, 2, 3, 4, 5$ denotes the number of components and $i = 1, 2, 3, 4, 5$ denotes the number of variables. The vectors w_{ik} are chosen such that the following conditions are met.

1. $\|w_k\| = 1$ for $k = 1, 2, 3, 4, 5$
2. Each of the $\mathcal{PC}_k$, maximises the variance $V\left\{w_k'\mathcal{PC}_k\right\}$ and
3. The covariance $COV\left\{w_k'\mathcal{PC}_k\, w_r{}'\mathcal{PC}_r\right\} = 0,\ \forall k < r$

The principal components are extracted from the linear combinations of the original variables that maximise the variance and have zero covariance with the previously extracted components. From a supervised perspective, extracted components can be viewed as new variables and from an unsupervised perspective, they may be viewed as "clusters".

Cluster analysis is a method for grouping data according to some measures of similarity. If we assume k distinct groups in our dataset, each with a specified centroid, then for each of the vectors $j = 1, 2, \ldots p$, we can evaluate the distance from $\mathbf{v}_j \in \mathcal{PC}_k$ to the nearest centroid from the set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\}$ as

$$\mathcal{D}_j(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k) = \min_{1 \le l \le k} d\left(\mathbf{x}_l,\ \mathbf{v}_j\right) \tag{17}$$

where $d(.)$ is an adopted measure of distance, and the clustering objective will then be to minimise the sum of the distances from each of the data points in $\mathcal{PC}_k$ to the nearest centroid. Optimal partitioning of the dataset requires identification of k vectors $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_k^* \in \mathbb{R}^n$ that solve the continuous optimisation function in Equation 18.

$$\min_{\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \in \mathbb{R}^n} f(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{j=1}^{p} \mathcal{D}_j(\mathbf{x}_1, \ldots, \mathbf{x}_k) \tag{18}$$

The solution to Equation 18 relates to the partial derivatives in Equations 7 and 8. Minimisation of the distances depends on the initial values in $\mathcal{C}$, hence if we let $z_{i=1,2,\ldots,n}$ be an indicator variable denoting group membership with unknown values, the search for the optimal solution can be through iterative smoothing of the random vector $x|(z = k)$, for which we can compute $\bar{\mu} = \mathbf{E}(x)$ and $\delta = \{\mu_k - \bar{\mu}|y = k \in \mathbf{c}_z\}$. In a labelled data scenario, $\{x_i, y_i\}$ $i = 1, 2, \ldots n$, Equation 18 transforms to the minimisation of Equation 19

$$f(\theta) = \sum_{i=1}^{n} \left[y_i - g(x_i;\ \theta)\right]^2 \tag{19}$$

where x_i are described by the parameters $\{\bar{\mu}$ and $\delta\} \in \theta$ and $g(x_i;\ \theta)$ are fitted values. Equations 17 to 19 relate to the K–Means clustering algorithm, which searches for clusters in numeric data based on a prespecified number of centroids. The decision on the initial number of centroids ultimately impinges on the detected clusters, and we attempt to address this issue via the two algorithms, below. Addressing spatio-temporal variations in dataset appeals naturally to dealing with randomness in data (Mwitondi and Said, 2013) and adopting interdisciplinary approaches to gain a unified understanding and interpretation of data modelling. In the next exposition, we present two algorithms: EstiMax and the Sample-Measure-Assess (SMA) algorithm (Mwitondi and Zargari, 2018), developed to address variations in data due to inherent randomness.

2.2.4 EstiMax Algorithm

Algorithm 1 below, adapted from a previous similar work (Mwitondi *et al.*, 2018a), is designed to identify naturally arising structures involving the five pollutants in Table 2 and the associated attributes–Times, DayTimes and Period. It searches for smoothing parameters that optimise the densities in Figure 5, while minimising the effect of randomness (Mwitondi, Moustafa and Hadi, 2013).

Algorithm 1

procedure EstiMax
  Set $\mathbf{X} = \left[x_{i,j}\right]$: Pollutants data
  Set z_{ik} as in Equation 11
  Set clusters $c_{z|x,j} \subset \mathbf{X}$; $j = 1 : K \ge 2$
  Initialise iteration: $m := 0$
  Sequential samples: $s_\kappa \subset \mathbf{X}$ where $\kappa = 1, 2, 3, \ldots, \lambda$
  Initialise: $\Theta_m\{.\} \leftarrow \theta_m \leftarrow c_{z|x,j}\{$Initial parameters for drawn samples$\}$ as in Equation 13
  for $\kappa = 1 : \lambda$ (*Large*) **do**
    while $m \le M$(Large) **do**
      Fit $c\left(z_{m.s_\kappa}|x_{m.s_\kappa}\right) \propto \frac{\pi_{m,j} f_{m,j}(x)}{\sum_{j=1}^{k} \pi_{m,j} f_{m,j}(x)} \leftarrow \hat{c}_{m.z|x_{m,j}\, s_m}$ as in Equation 12
      Update $m := m + 1$
      Update $\Theta_m\{.\} \leftarrow \theta \leftarrow \hat{c}_{m.y|x_{m,j}\, s_m}$ as in Equation 13 through 15
    end while
    Update $\kappa := \kappa + 1$
  end for
end procedure

The EstiMaxi algorithm provides general mechanics for estimating crucial parameters of the pollutants and their likelihoods. Our interest is to obtain multiple sets of parameters $\Theta\{.\}$ in the sampled periods for accurate and consistent estimation. We can then associate the parameters across samples with the time-related variables in Table 2. The SMA algorithm (Mwitondi and Said, 2021) below, invokes the EstiMax, performs comparative analysis and determines the best model. However, this sequential relationship is not a requirement for implementing either of the algorithms, because the EstiMax can estimate parameters from any input dataset, while the SMA can be executed with any predetermined parameters.

Algorithm 2

procedure SMA
  Set $\mathbf{X} = \left[x_{i,j}\right]$: Pollutant datasets
  Set $c_k : k = 1, 2, ..., K$ initial clusters
  Call **Algorithm 1**
  Output Performance Parameters in $\Theta_m\{.\}$
  Initialise: $\Pi_{cp} := \Pi_{cp}(.)$: Comparative parameters
  Learn $F(\phi) = \underset{x,y\sim D}{(P)}\left[\phi(x) \neq y\right]$ based on a chosen learning model
  for $i := 1 \rightarrow \varkappa$ **do**: Set $\varkappa$ large and search for optimal values iteratively
    while $s \leq 50\%$ of $\left[x_{\nu,\tau}\right]$ **do** Vary sample sizes to up to the nearest integer 50% of X
      Sampling for Training: $s_{tr} \leftarrow X$
      Sampling for Testing: $s_{ts} \leftarrow X$
      Fit Training and Testing Models $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ with current parameters
      Update Training Parameters: $\Theta_{tr}(.) \leftarrow \Theta_{tr}$
      Update Testing Parameters: $\Theta_{ts}(.) \leftarrow \Theta_{ts}$
      Compare: $\Phi(.)_{tr}$ with $\Phi(.)_{ts}$: Using visual plots and/or animation
      Update Comparative Parameters: $\Pi(.)_{cp} \leftarrow \Phi(.)_{tr,ts}$
      Assess: $P\left(\Psi_{D,POP} \geq \Psi_{B,POP}\right) = 1 \Longleftrightarrow \mathbb{E}\left[\Psi_{D,POP} - \Psi_{B,POP}\right] = \mathbb{E}\left[\Delta\right] \geq 0$
    end while
  end for
  Output the Best Models $\hat{\mathcal{L}}_{tr,ts}$ based on $\mathbb{E}\left[\Delta\right] \geq 0$
end procedure

The population error will always be greater than or equal to the training sample error drawn from the same population, hence the expected difference is guaranteed to be greater than or equal to 0 ($\mathbb{E}\left[\Delta\right] \geq 0$). Section 3 focuses on exploratory data analyses, optimal estimation of the key parameters for the pollutants in Table 2, and their spatio-temporal associations, thus providing insights into assessing pollution patterns and relevant dynamics in southern China. The modelling techniques adopted—PCA, K-Means and the EM algorithm—are based on their underlying mechanics, particularly the influence of the starting point for the last two and their adaptability to the two algorithms.

3 IMPLEMENTATION

Application of the methods to the selected datasets is predicated on the premises that variations in data sources, quality and interpretations entail a good understanding of the overall behaviour of data, as a crucial performance indicator. This section covers exploratory data analysis, identifies naturally arising structures in the sampled data and builds associations among different pollutants over different time periods across the collected samples.

3.1 EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a standard method that is usually deployed ahead of data modelling to provide useful insights into understanding data. It delivers key characteristics of targeted datasets, often through numerical and visual summaries (Komorowski *et al.*, 2016). It helps in revealing the overall data behaviour, in detecting patterns, and in spotting anomalies (Ridzuan and Zainon, 2022), as well as being useful in checking the validity of assumptions. Most importantly, its results are useful guidelines to evaluate the appropriateness of modelling techniques.

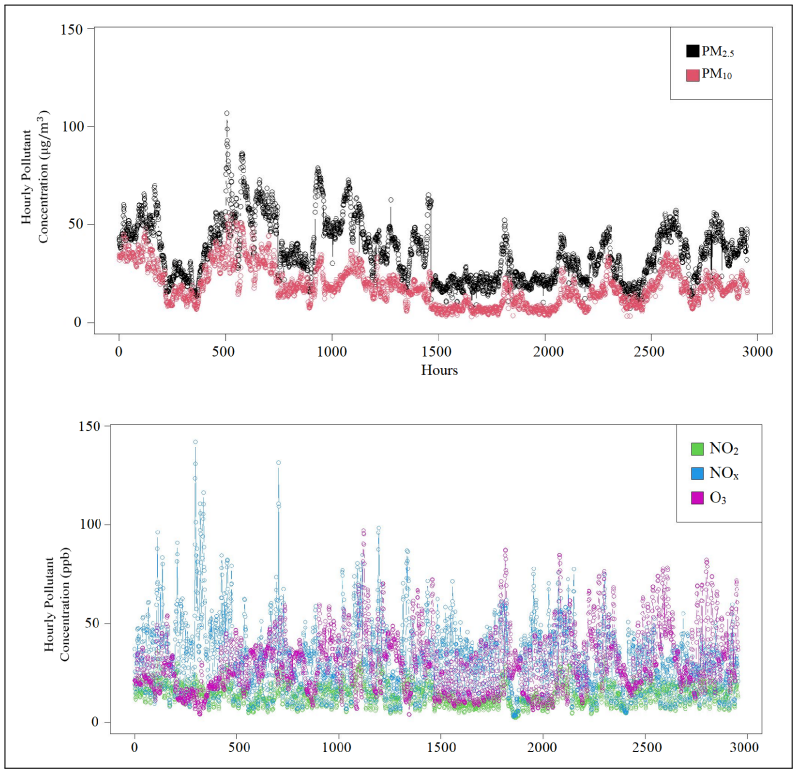

Figure 2 Hourly averaged concentrations for all sampled pollutants.

The left-hand side panel of Figure 2 shows hourly time series of $PM_{2.5}$ and PM_{10} concentrations, during the investigated time period, measured in micrograms per cubic metre ($\mu g/m^3$). The right-hand side panel are for NO_2, NO_X and O_3, measured in "parts per billion" (ppb)–a concentration unit for the number of parts of a substance per billion parts. In both cases, the horizontal axis displays the study period in hours. The time series shows pollution patterns across pollutants. For instance, they generally tend to be lower during the summer period and higher during winter –towards the beginning and end of year. Both NO_X and $PM_{2.5}$ have the highest hourly pollution levels during the first 700 hours of the year and lowest during the last half of the year. Further, NO_2 pollution appears to be within the same margins throughout the year, whereas O_3 tends to peak at regular intervals after the first third of the year. We can relate these patterns to variables of interest such as **DayTimes** and **Period**, to assess the time effect of pollutants.

START HOUR	END HOUR	HOURS RANGE	CATEGORIES OF DAY
00:00hrs 01st-Jan-2023	23:00hrs 31st-Jan-2023	1st – 744th hour	Night, Morning, Day, Evening
00:00hrs 01st-Apr-2023	23:00hrs 30th-Apr-2023	745th – 1464th hour	Night, Morning, Day, Evening
00:00hrs 01st-Jul-2023	23:00hrs 31st-Jul-2023	1465st – 2208th hour	Night, Morning, Day, Evening
00:00hrs 01st-Oct-2023	23:00hrs 31st-Oct-2023	2209th – 2952th hour	Night, Morning, Day, Evening

Table 3 Daily and monthly averages.

The time series plots in Figure 2 span across a four month period—January, April, July and October—each further discretised into a repeating sequence of "Night, Morning, Day, Evening" (in Table 3), with corresponding "Start" and "End" hours as shown in Table 2. These trends can also be visualised based on daily averages, in which case patterns representing categories of day, i.e., night, morning, day and evening, can be seen (see Table 4). Steps 8 to 18 of **Algorithm # 2** can be applied, with appropriate parameters, to sample through the data for other choices of interest, e.g., day and night or half months. The plots of pollution trends for the four months were fairly consistent with Figure 2.

3.2 UNDERSTANDING DISTRIBUTIONAL BEHAVIOUR

Table 4 exhibits day and annual average pollution levels, where PM_{10}, $PM_{2.5}$ and NO_2 are lowest in July. The monthly average PM_{10} concentration is 19.26 $\mu g/m^3$ (with a standard deviation of 7.91 $\mu g/m^3$), and the corresponding daily average is 20.29 $\mu g/m^3$ (with a standard deviation

AVERAGE TIME PERIODS	PM_{10} $\mu g/m^3$	NO_2 ppb	$NO_x = NO + NO_2$ ppb	O_3 ppb	$PM_{2.5}$ $\mu g/m^3$
Day	20.08	15.77	38.00	39.51	35.52
Evening	21.51	15.99	35.85	30.29	38.41
Morning	19.96	13.61	34.79	23.70	34.68
Night	19.60	11.05	22.37	25.47	35.35
January	28.36	15.30	39.38	26.06	45.01
April	19.51	15.93	33.47	34.17	41.77
July	9.05	11.69	35.56	29.44	23.05
October	20.14	14.49	29.79	35.79	33.84

Table 4 Daily and monthly average concentrations of each pollutant.

of 0.84 $\mu g/m^3$). The corresponding monthly and daily averages (deviations) for NO_2 are 14.35 (1.86) ppb and 14.11 (2.30) ppb respectively. For NO_X, the corresponding values are 34.56 (4.00) ppb and 32.75 (7.04) ppb respectively, whereas for O_3 and $PM_{2.5}$, the corresponding averages and variations are 31.37 (4.44) ppb, 29.74 (7.08) ppb, 35.92 (9.78) ppb and 35.99 (1.65) ppb respectively. The periodic averages in Table 4 and the variations stated above, provide a summary of the pollutants' distributional behaviour.

Over time, the parameter variations in Table 4 can be used to guide optimisation models for pollution monitoring. Our strategy is to apply algorithms 1 and 2 to estimate, maximise and compare key parameters of pollution. Figure 3 exhibits how central tendency and variation parameters can vary over time. Monthly averages (in red) and standard deviations (in blue) are shown in Figure 3a, while Figure 3b shows the equivalent daily averages.

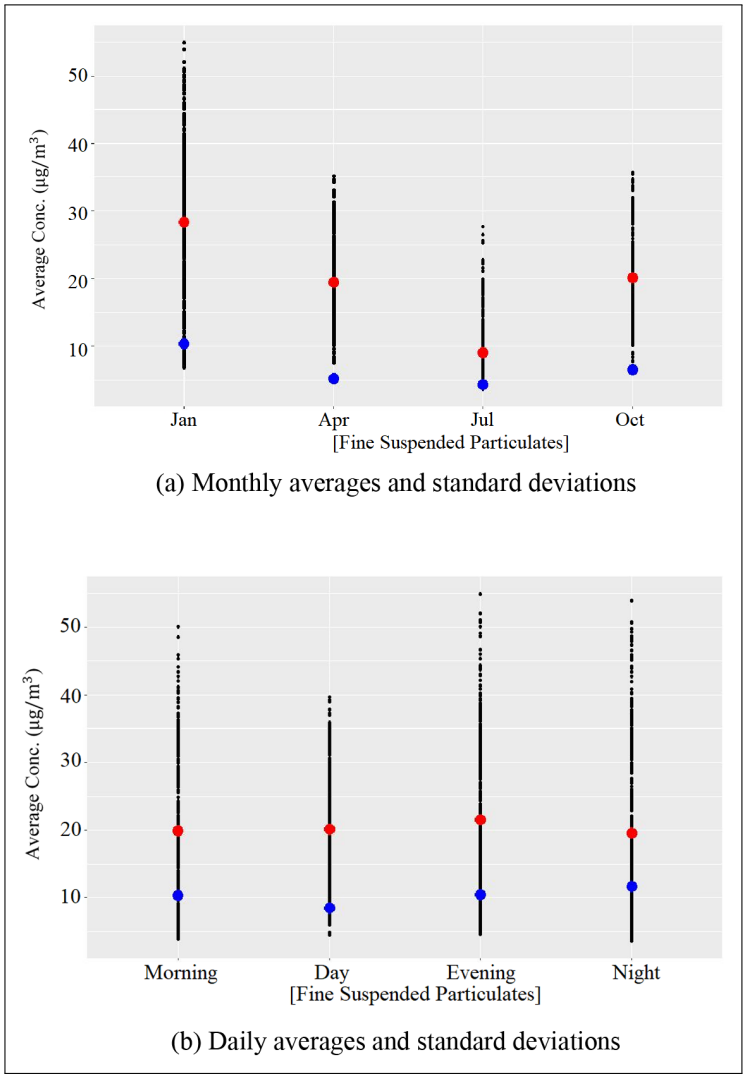

Figure 3 Fine suspended particulates monthly and daily averages and variations.

As an example of the time-related patterns of pollution, consider the distribution of O_3 levels over the year, as exhibited in Figure 4. It can be observed that O_3 attains its highest concentration during January for all four discretised time periods of day, and that the lowest concentration occurred in July 2023. The variation across different time periods within a day (particularly during July) when pollution levels drop before rising again, is of interest. Understanding the behaviour of such spatio-temporal variations is crucial to gain insights into ways of mitigating the potential impacts of air pollution on the well-being of our surrounding environment and neighbourhoods.

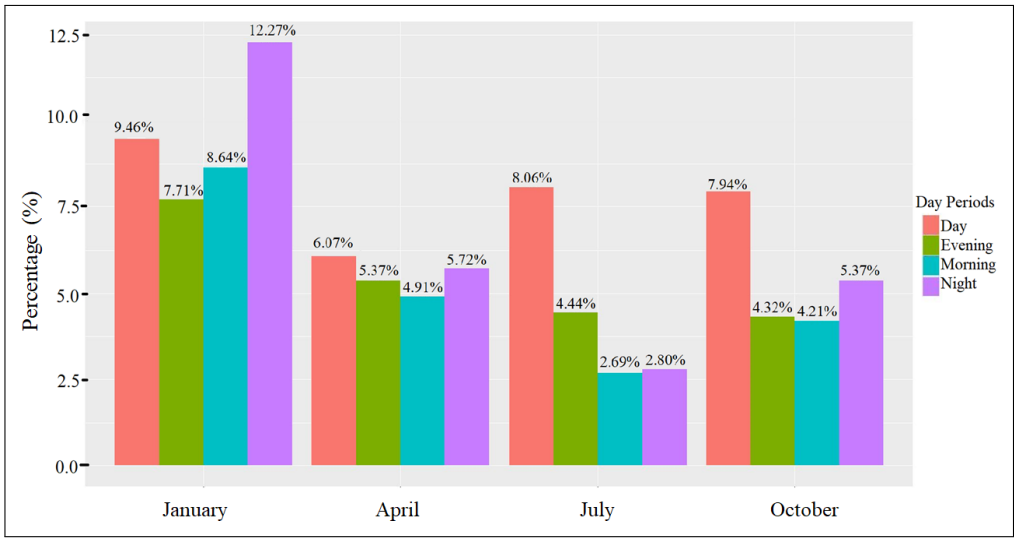

Figure 4 Pollution levels across the year 2023.

Figure 5 shows the density plots for the hourly averages across day time periods, computed across stations, of each of the five pollutants. The horizontal axes in Figure 5a and 5b are micrograms per cubic metre ($\mu g/m^3$) and ppb respectively. In both cases, each density is associated with free parameters that describe its structure, centrality and dispersion, i.e., group membership, mean and variation (π_k, μ_k, σ_k) respectively. These examples are based on datasets acquired from eight stations in southern China: Station-1 (CB_R), Station-2 (CL_R),

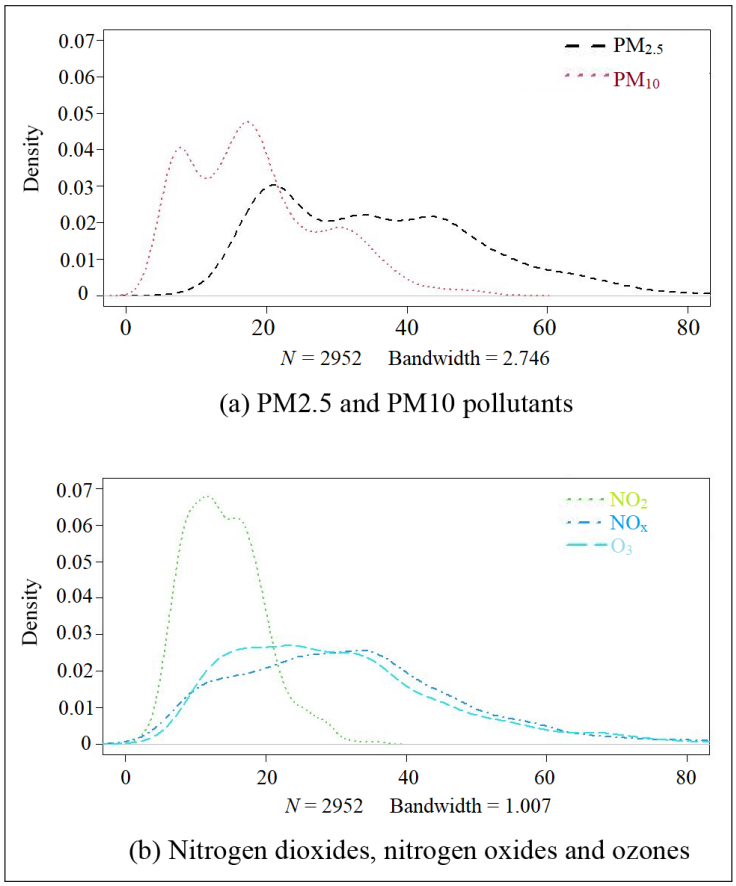

Figure 5 Density distribution of pollutants across day time periods.

Station-3 (XCNAQ245), Station-4 (XCNAQ246), Station-5 (XCNAQ250), Station-6 (XCNAQ252), Station-7 (XCNAQ253), and Station-8 (XCNAQ254).

The sampled stations are distributed in different parts within our investigated domain, thus serve as good representatives of pollution figures across the sampled region. It is important to note that the sensors used for data recording adopt different measurement techniques and therefore the densities in Figure 5 are not intended for direct comparison. However, their variation patterns can be used to gain useful insights into different pollutants on a typical day.

3.2.1 Optimisation of Parameters

Figure 6 shows a 2–dimensional clustering of raw pollution datasets, which allows readers to obtain a clear visualisation of the numeric data presented in Table 2. The data points are represented by points in the ellipsoid plots, using principal components or multidimensional scaling. We use the method to illustrate the optimisation process adopted in estimating θ, as defined in Equation 2. The plots represent a set of two, three and four clusters generated from average pollution levels of the five pollutants. Each of the three set of clusters is formed around a set of centroids that can be determined in a number of ways–such as random initialisation or data–dependent parameters.

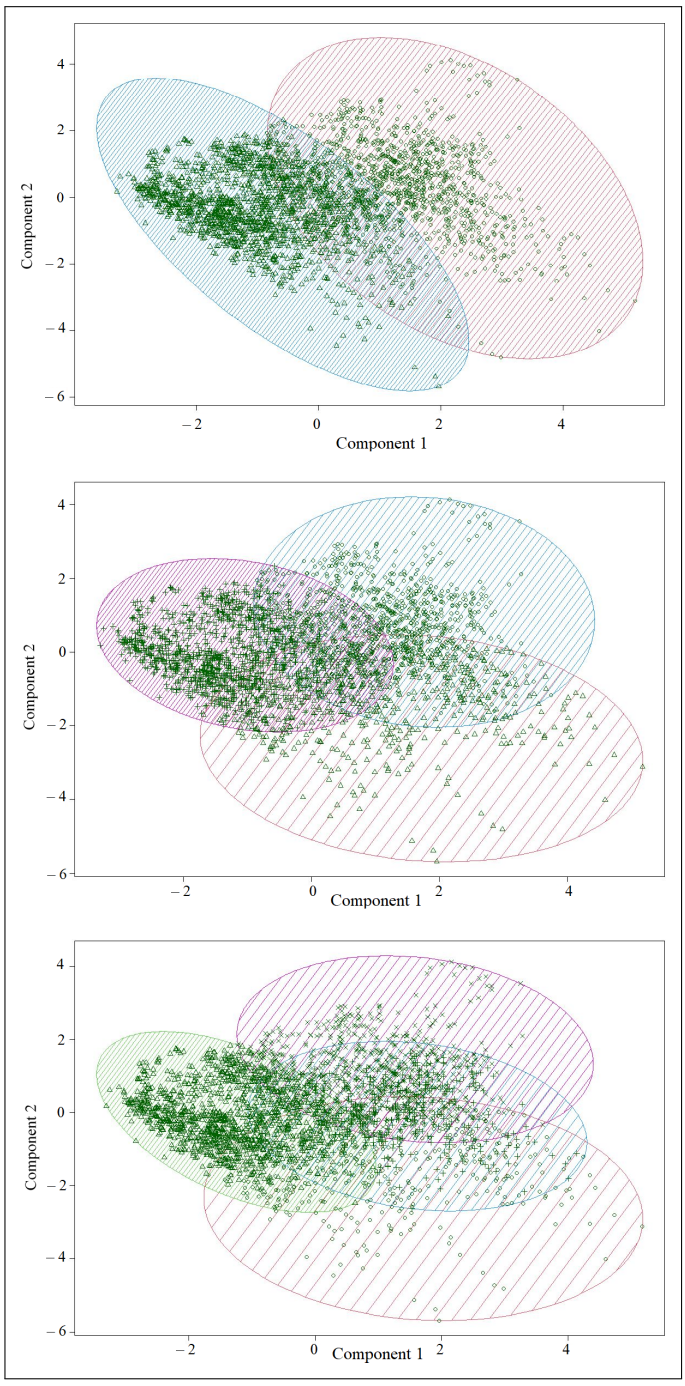

Figure 6 Pollution data points on a multidimensional scaling.

Table 5 exhibits the contributions of each pollutant into the formation of components, as well as the final centroids for each of the three sets of clusters. In all three cases, the two components contributed 76.41% of the statistical variation.

Table 5 Centroids of the selected clusters formed.

POLLUTANT	2 CLUSTER CENTRES	3 CLUSTER CENTRES	4 CLUSTER CENTRES
Fine Suspended Particulates (PM_{10})	• 12.11 • 25.85	• 11.79 • 21.44 • 25.55	• 11.53 • 20.38 • 20.48 • 28.54
Nitrogen Dioxide (NO_2)	• 11.79 • 16.70	• 10.93 • 20.61 • 15.07	• 10.88 • 20.26 • 17.99 • 13.56
Ozones (O_3)	• 22.12 • 39.53	• 22.49 • 26.54 • 41.50	• 22.51 • 23.37 • 55.29 • 30.69
Nitrogen Oxides ($NO_x = NO + NO_2$)	• 29.77 • 34.85	• 25.51 • 58.79 • 28.43	• 25.44 • 60.33 • 35.99 • 25.34
Respirable Suspended Particulates ($PM_{2.5}$)	• 25.15 • 48.14	• 24.85 • 37.94 • 48.59	• 24.42 • 36.08 • 40.39 • 53.02

An illustration of estimation and maximisation of key parameters of the pollutants using the EstiMax algorithm is given in Figure 7a and 7b based on raw PM_{10} and $PM_{2.5}$ concentrations from the 8 stations in southern China. In each case, the dotted thin lines represent the average across samples and the thick line is the maximised estimated average across the samples. The algorithm does quite well in capturing the multi-modality of the pollutants.

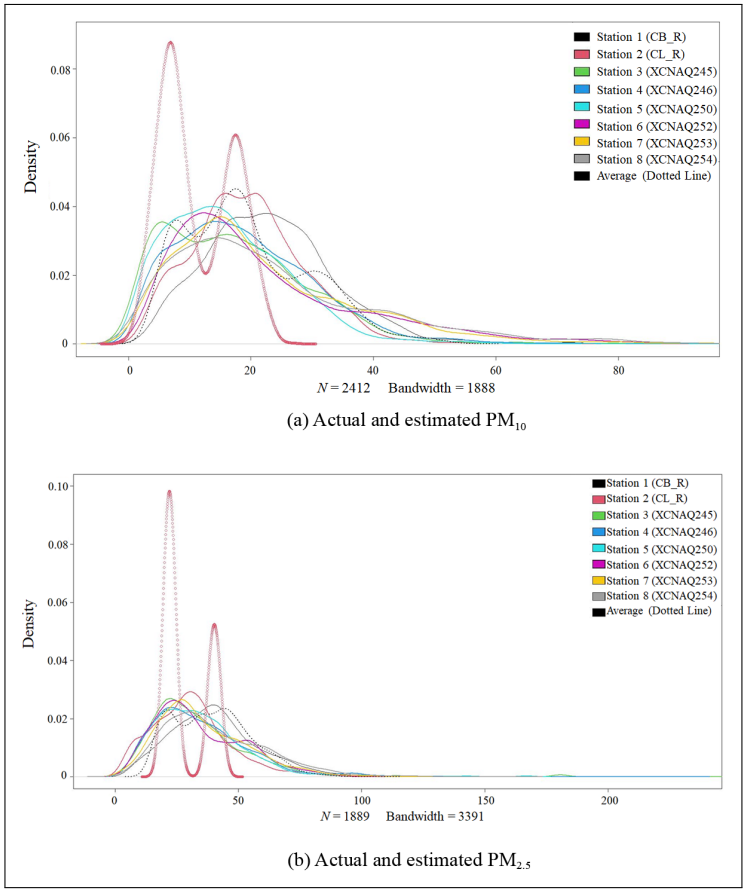

Figure 7 Actual and estimated PM_{10} (Left) and $PM_{2.5}$ (Right).

The bimodality of the estimated pollution levels across the eight stations were the best results after comparing 2, 3, 4 and 5 modes, in relation to the raw patterns. The bimodality, effectively, implies that there are two pollution levels across the eight monitoring stations: high and low, although there are clear outliers in the upper tails in each case.

3.2.2 Extracting Components

If we denote the overall state of pollution by $\mathcal{P}$, our interest lies on the multicollinearity of the known factors, which cause inaccurate estimates of various parameters, and as a result affect future modelled outputs or forecasts. One way of addressing this issue is to apply PCA –a statistical method that reduces data dimension, and maximises variation within the same dataset. Ultimately, this results in the estimation of parameters (θ) in Algorithm 1, which improves the accuracy of $\mathcal{P}$ for being used as prior information for ongoing pollution monitoring within a prescribed spatial domain.

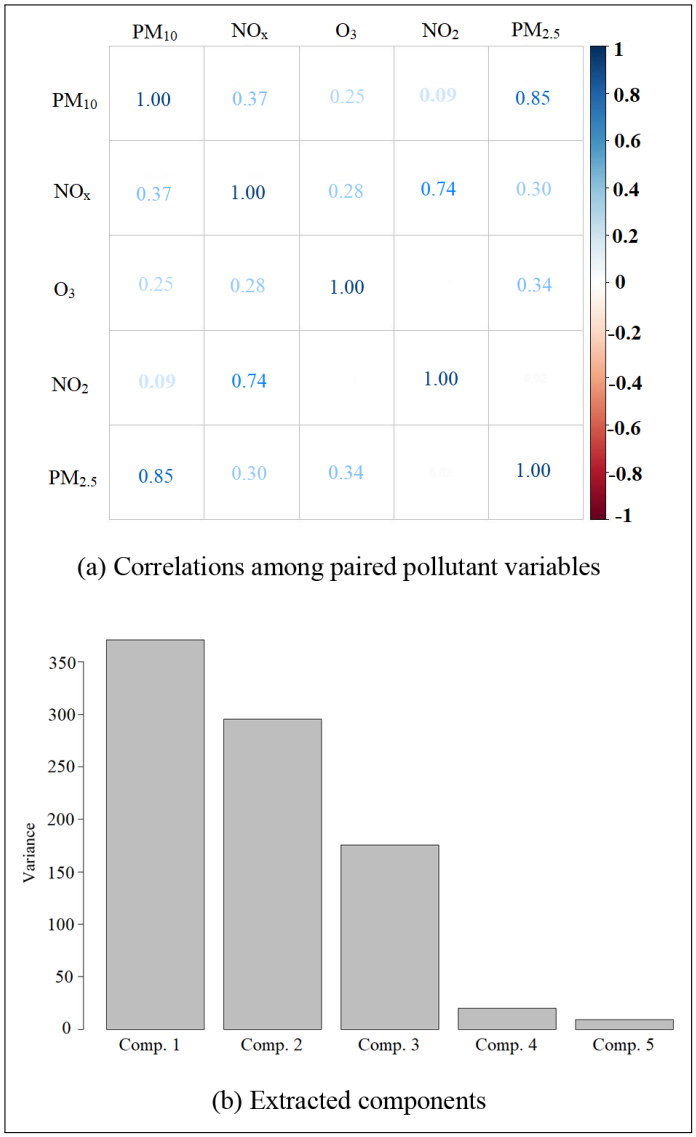

Figure 8 Correlations among paired pollutant variables.

Figure 8a shows the correlation figures among each pair of the five pollutant variables, while Figure 8b shows the extracted components. In this case, the standard deviations in Component 1 through Component 5 are 19.26, 17.18, 13.24, 4.46 and 3.01, with proportions of variance equal to 0.426, 0.339, 0.201, 0.0228 and 0.010 respectively. Note that the first two components account for 76.5% of the variation in the pollution dataset.

Table 6 shows the contribution of variables in each component. It can be seen that PM_{10} has a very strong relationship with component 4, while the absence of Nitrogen Oxide in Component 2 is most pronounced, as is that of Nitrogen Dioxide in Component 5. Similarly, the absence of ozones in Component 3 is seemingly vital in its formation.

	LOADINGS				
POLLUTANT	**COMPONENT 1**	**COMPONENT 2**	**COMPONENT 3**	**COMPONENT 4**	**COMPONENT 5**
PM_{10}	0.399		0.334	0.822	0.221
Nitrogen Dioxide (NO_2)	0.162	–0.200		0.203	–0.944
Ozones (O_3)	0.536	0.199	–0.816		
Nitrogen Oxide (NO_x)	0.258	–0.935			0.225
$PM_{2.5}$	0.679	0.203	0.469	–0.524	

Table 6 Component loadings–contribution of each pollutant in each component.

For a clearer interpretation of the numerical and graphical results in Section 3, we need to relate them to the modelling strategy in Section 2.2–particularly to Algorithms 1 and 2. Part of the modelling strategy provides the theoretical foundations on which we can understand the overall data behaviour, and part of it provides aspects of its practical implementation. The main idea is to attain optimal interpretability of modelling results which, typically, derive from training and validating statistical models on "random samples" and ultimately applying them on new data that is also "random", entailing variations across samples. Note that, mathematically, from an $m \times p$ dimensional dataset, a total of p components can be extracted, but, usually only a few will explain the variation in the data. It is the foregoing data variation that Algorithms 1 and 2 seek to address. For instance, the centroids of the formed clusters in Table 6 are based on 1, 2, 3, 4 and 5 clusters, although for PM_{10} and Ozone, the maximum of 75 and 114 respectively could have been extracted. Drawing new samples of PM_{10} and Ozone from similar data sources isn't necessarily going to generate the same number of components or the same loadings. The five clusters in Table 6 are judged optimal according to the repeated sampling via Algorithms 1 and 2. The same applies to correspondence analysis.

3.2.3 Correspondence Analysis

The average pollution levels for each of the five pollutants from each of the eight locations are given as continuous data. Their density plots exhibit multi-modality –a distributional behaviour that is supportive of discretisation of each variable. If we denote the average vector by $\mathcal{V}_a(x)$, we can discretise it by setting the following rule:

$$\mathcal{V}_a(x) = \begin{cases} \textbf{if } x \leq \text{Lower Estimated Mean} & \textbf{Low} \\ \textbf{if } x \leq \text{Mid Estimated Mean} & \textbf{Medium} \\ \textbf{else if } x > \text{Higher Estimated Mean} & \textbf{High} \end{cases} \tag{20}$$

Equation 20 describes how an ordered continuous data vector can be visualised based on its overall behaviour. Mean estimates are in accordance with the EstiMax algorithm, and initial points can range from completely random to data-dependent, such as those in Figure 3 or other data parameters, such as using percentiles or quartiles. In each case, a quantitative vector is broken down into discrete segments that form categories of visual analysis based on correspondence analysis. The technique enables the visualisation of associations between different categories of selected data attributes in a 2–dimensional space. It seeks to establish associations between some row elements and some column elements, generating orthogonal components, with maximisation of variation in the data in mind.

Figure 9a represents associations between PM_{10}, discretised averages with the Time of Day, while Figure 9b shows the same association with annual periods. In both cases, the axes measure the levels of variation in the data, with the extreme left of the horizontal axis representing the negative measure and the extreme right representing the most positive measure. Similar definitions are applicable for the south and north directions of the vertical axis. The highest variation in both panels is accounted for by the first component, at 95.7% and 92.7% respectively.

Tables 7 and 8 represent the frequencies of each of the three categories during the day and the year 2023, respectively. Note how the High and Medium categories dominate in each case. It is worth mentioning that this distribution is, and should always be, conditional on the distributional behaviour of the discretised variable, since the number of classes does affect how

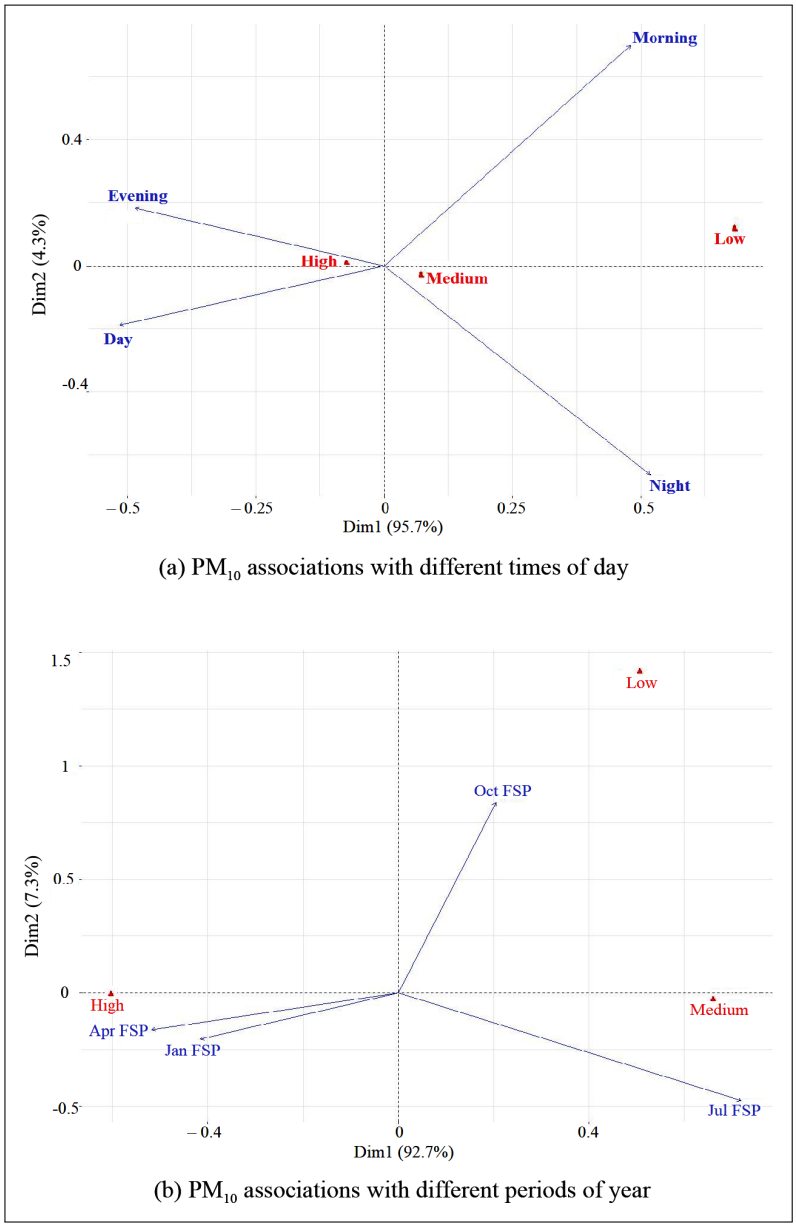

Figure 9 PM_{10} associations with daily and annual periods.

many cases fall into each class. Our interest in Figure 9 is to find out the points that contribute to the solution provided by the method which, in this case, presents associations between the pollutants $PM_{2.5}$ and PM_{10} with different time periods. CA forms these patterns based on "expected values"–associating row and column labels for the disparity between "expected" and "observed" values, in order to explain the percentage of variance in the data.

	HIGH	LOW	MEDIUM
Day	484	4	373
Evening	352	3	260
Morning	369	20	349
Night	351	17	370

Table 7 Row points vs Principal Dimension 1.

	HIGH	LOW	MEDIUM
January	587	0	157
April	618	0	102
July	58	4	682
October	293	40	411

Table 8 Columns vs Principal Dimension 1.

Figure 9a shows the relationship between PM_{10} and times of day. The strength of the relationship is measured by the distance between two points and the tightness of the angle: the tighter the angle, the stronger the relationship, e.g., the angles formed between "Day" and "Evening" with "High" pollution. Orthogonal angles (90°) indicate no relationship, while a 180° angle indicates negative association. The length of the line connecting the row label to the origin indicates the strength of the row label association, e.g., "Morning" and "Night" with respect to "Medium" and "Low" pollution levels. Being farther away from the origin means more closely associated with the factors in the proximity. In Figure 9a and 9b, most time periods are farther from the origin, without being very close to any factor.

Table 9 presents the row points, most associated with the first principal dimensions (PD) for the association between PM_{10} and Day Times, while Table 10 presents similar information for PM_{10} with annual periods. Evening times and January have the lowest row points association with the second PD, whereas Night Times and October have the highest association with the second dimension. These are quite interesting patterns, worth following through as they potentially indicate commonalities between those periods which may relate to specific activities or lack of them.

	DIMENSION 1	DIMENSION 2
Day	26.50596	3.576712
Evening	23.52921	3.377988
Morning	23.04545	49.011043
Night	26.91938	44.034257

Table 9 Row points vs PD 1 for Day Times.

	DIMENSION 1	DIMENSION 2
April	26.778387	2.672733
January	17.240575	4.196079
July	51.795246	22.524998
October	4.185793	70.606190

Table 10 Columns vs PD 1 for Annual Periods.

A further comparative analysis from eight different stations is provided in Figure 10, in which PM_{10} and $PM_{2.5}$ are compared for stations 1 and 8, positioned further apart from each other. The top two panels present $PM_{2.5}$ concentrations, while the bottom panel presents PM_{10} concentrations within the same stations. In both cases, variations within the four months of 2023 are more obvious than variations within different periods of a day.

The discretisation in Equation 20 provides the comparative basis in line 15 of the SMA algorithm (Algorithm 2), which can be visualised via plotting. On the basis of such comparison, one can update parameters within the algorithm, and conduct corresponding assessment of $\mathbb{E}[\Delta]$, to determine the difference between predicted and actual parameters.

4 CONCLUDING REMARKS

This paper proposed a data-driven robust method for addressing societal challenges, focusing on air pollution –a problem that cuts across disciplines, sectors and geographic borders. The focus on air pollution, SDGs and southern China was inspired by the human–nature interactions and their mutual impact, as we know them, i.e., air pollution affecting and/or being affected by different SDGs. The link between air pollution and various aspects of SDGs, the need for developing a replicable model and the choice of the study area were used to highlight the purpose of the study. Focusing on southern China was motivated by the country's global role as a polluter and mitigator of pollution, while the focus on "robustness" was motivated by the need for developing models that could be replicated in a spatio-temporal context. The main idea was to use the "most parsimonious model" to highlight potential data–driven solutions to more complex regional and global scenarios. It was assumed that gaining insights into

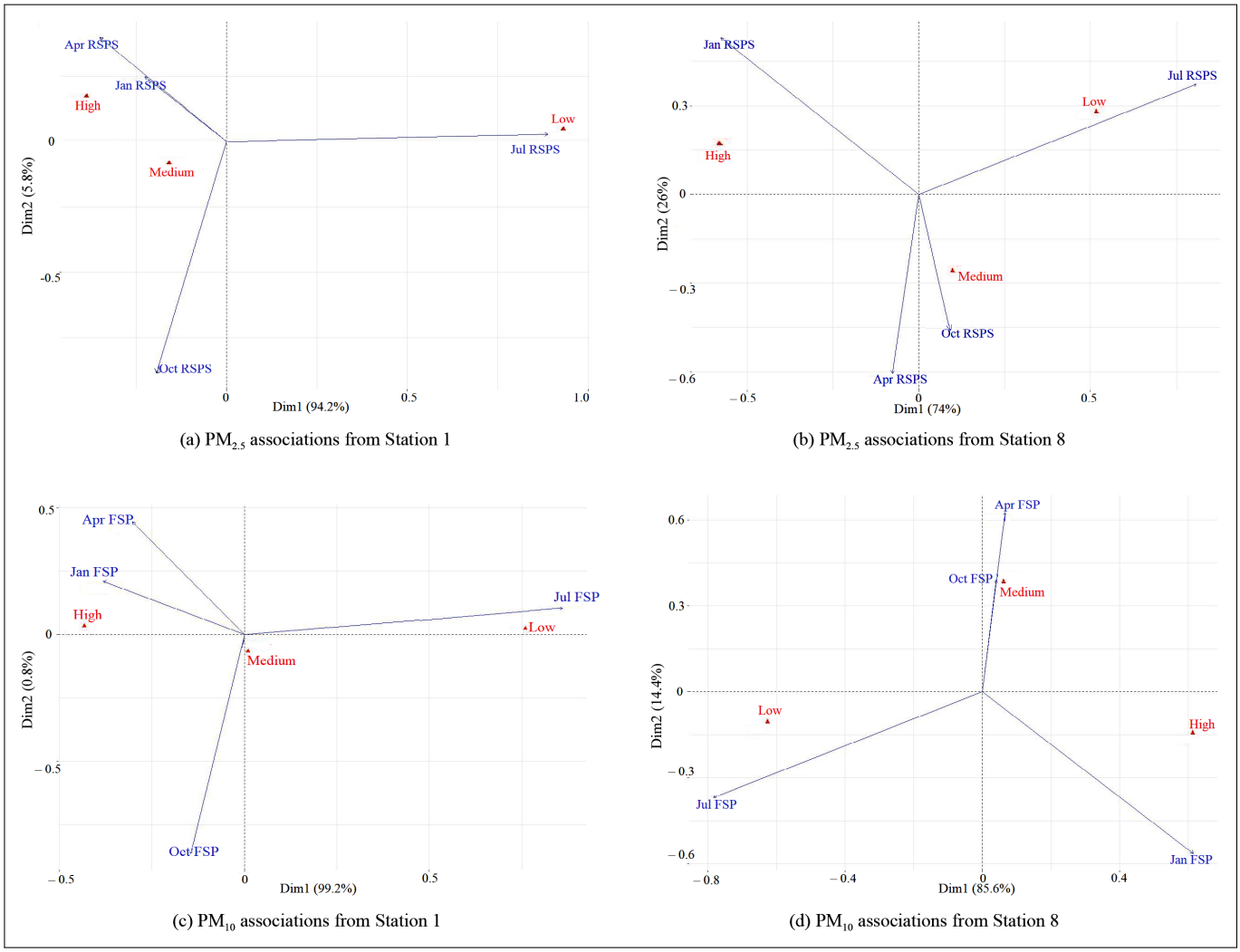

Figure 10 PM_{10} and $PM_{2.5}$ associations from Stations 1 and 8.

spatio-temporal patterns of the common pollutants in southern China would highlight the path towards mitigating the impact of pollution elsewhere. It was further assumed that the overall pollution levels in southern China were mainly contributed by the five major pollutants within the region, and that these five attributes could be used as predictors of the pollution levels in the region and, potentially, elsewhere.

The paper investigated associations among common air pollutants in southern China during specific months in 2023 and at different times of a day. Its key motivation was twofold. Firstly, human and natural activities constantly generate multifaceted pollution data much faster than our ability to process them and, secondly, most environmental modelling approaches still face model optimisation challenges that derive from inherent randomness in data (Mwitondi, Munyakazi and Gatsheni, 2018b). Hence, developing tools and methods for harnessing, processing and sharing such data is crucial to address a wide range of societal challenges, such as human health, food security and other SDG-related challenges and opportunities.

Two algorithms were used to maximise and optimise air quality parameters that best describe associations between different attributes of interest, e.g., between specific pollutants and timelines; they provided insights into policy formulation on spatio-temporal mitigation of air pollution. They also generated static, interactive and easily interpretable visual outputs, that are pivotal in optimising operational efficiency. For instance, the parameters in Figure 3 relate to monthly and daily averaged PM_{10} concentration figures, which are subject to spatio-temporal variation. Monitoring such variation within and across samples is crucial to improve our understanding of how a specific spatial domain could be affected by pollution that occurs in the lower atmosphere. Further, the algorithms used unobservable (latent) variables to reduce data dimensionality, and combined observable variables to make them more interpretable. While the algorithms balance the power of data, machine learning techniques and underlying domain knowledge, they also contribute towards attaining mutual understanding across fields and sectors. Overall, the current study has implemented rigorous interdisciplinary approaches

in combining sophisticated data science algorithms and relevant domain knowledge of our atmosphere.

A comparative analysis with previous studies was based on model optimisation, which is what the two algorithms sought to achieve. The novelty of the study hinges on addressing data randomness and on highlighting the path to addressing the 'non-orthogonality' of SDGs. The two algorithms were used to generate optimal associations of temporal patterns (within different hours of the same day, monthly or annually) with relevant pollutant attributes; they provided particularly useful insights into our understanding of the overall state of pollution within the atmosphere of the southern China region, Hong Kong and Macau. Balancing the power of data, machine learning techniques and underlying domain knowledge through the two algorithms exhibited superiority over standard machine learning applications. Identifying such associations aligns with the key aspects of SDGs, particularly their "non-orthogonality", and it can help to highlight paths to a unified and interdisciplinary understanding of the triggers of the SDG indicators.

The two general and seven specific objectives in Section 1.3 were fully met, except objective #2 (b), which remains a subject for further research. However, the findings of this research will contribute to a better understanding of how to deal with the challenges posed by the non-orthogonality of socioeconomic, technical, and environmental attributes of the SDGs. That is because the non-orthogonality of SDGs masks a lot of potentially useful data that researchers need to dig up and share publicly with the wider scientific community, as well as with policymakers. For that, there is no better testing ground on how to build analytical bridges across disciplines and sectors than on the SDG spectrum. This work is expected to highlight novel directions into the environmental and other aspects of SDG modeling. It is worth noting that applications of the proposed techniques are not confined to air pollution. The two algorithms are readily adaptable to anomaly detection for use in different applications in industry, medicine and business (Li and Jung, 2023; Li *et al.*, 2023). In this application, however, rather than just uncovering "abnormal" events, we mapped them to periodic patterns and, hence, provided better insights to stakeholders, which potentially connects to and influences policymakers.

ACKNOWLEDGEMENTS

We would like to thank our respective institutions – Qatar University, SESRI Institute; The Chinese University of Hong Kong, Department of Mathematics; and The Hong Kong University of Science and Technology, Department of Mathematics, for allowing us time to complete this manuscript. We are also deeply grateful to all those who reviewed the manuscript at different stages of its development, and who provided very constructive suggestions, most of which we adopted. Last, but not least, this work would not have been completed without the patience and support of many colleagues, friends and family members whose daily lives were touched by the development of this work.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Kassim Mwitondi orcid.org/0000-0003-1134-547X
Qatar University, SESRI Institute, Qatar

Hugo Wai Leung Mak orcid.org/0000-0002-7033-6218
Dept of Mathematics, The Chinese University of Hong Kong, Hong Kong, China; Dept of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

REFERENCES

Anser, M.K., Ali, S., Mansoor, A., ur Rahman, S., Lodhi, M.S., Naseem, I. and **Zaman, K.** (2024) 'Deciphering the dynamics of human-environment interaction in China: Insights into renewable energy, sustainable consumption patterns, and carbon emissions', *Sustainable Futures*, 7, p. 100184. Available at: https://doi.org/10.1016/j.sftr.2024.100184

Cai, W., Xu, X., Cheng, X., Wei, F., Qiu, X. and **Zhu, W.** (2020) 'Impact of "blocking" structure in the troposphere on the wintertime persistent heavy air pollution in northern China', *Science of The Total Environment*, 741, p. 140325. Available at: https://doi.org/10.1016/j.scitotenv.2020.140325

Chapmann, J. (2017) *Machine Learning: Fundamental Algorithms for Supervised and Unsupervised Learning With Real-World Applications (Advanced Data Analytics)*. CreateSpace Independent Publishing Platform.

Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C., Grujic, V., Model, T., Kubinec, R. and **Barceló, J.** (2024) 'A general primer for data harmonization', *Scientific data*, 11(1), p. 152. Available at: https://doi.org/10.1038/s41597-024-02956-3

Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., Yang, J. and **Tao, J.** (2022) 'Machine learning-based estimation of ground-level NO_2 concentrations over China', *Science of the Total Environment*, 807, p. 150721. Available at: https://doi.org/10.1016/j.scitotenv.2021.150721

Dempster, A.P., Laird, N.M. and **Rubin, D.B.** (1977) 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), pp. 1–22. Available at: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Diaz-de Arcaya, J., Garcia-Perez, A., Bonilla, L., Miñón, R. and **Torre-Bastida, A.I.** (2025) 'Data harmonization as a keystone for data spaces: Challenges, techniques, and future trends', in *2025 10th International Conference on Smart and Sustainable Technologies (SpliTech)*, IEEE, pp. 1–6. Available at: https://doi.org/10.23919/SpliTech65624.2025.11091719

Du, J., Qiao, F., Lu, P. and **Yu, L.** (2022) 'Forecasting ground-level ozone concentration levels using machine learning', *Resources, Conservation and Recycling*, 184, p. 106380. Available at: https://doi.org/10.1016/j.resconrec.2022.106380

Edo, G.I., Itoje-akpokiniovo, L.O., Obasohan, P., Ikpekoro, V.O., Samuel, P.O., Jikah, A.N., Nosu, L.C., Ekokotu, H.A., Ugbune, U., Oghroro, E.E.A. *et al.* (2024) 'Impact of environmental pollution from human activities on water, air quality and climate change', *Ecological Frontiers* 44(5), pp. 874–889. Available at: https://doi.org/10.1016/j.ecofro.2024.02.014

Feng, H., Zou, B., Wang, J. and **Gu, X.** (2019) 'Dominant variables of global air pollution-climate interaction: Geographic insight', *Ecological Indicators*, 99, pp. 251–260. Available at: https://doi.org/10.1016/j.ecolind.2018.12.038

Feng, X., Wei, S. and **Wang, S.** (2020) 'Temperature inversions in the atmospheric boundary layer and lower troposphere over the Sichuan Basin, China: Climatology and impacts on air pollution', *Science of the Total Environment*, 726, p. 138579. Available at: https://doi.org/10.1016/j.scitotenv.2020.138579

Global Alliance on Health and Pollution (2023) 'Global alliance on health and pollution'. Available at: https://www.gahp.org/

Goncharenko, L.P., Harvey, V.L., Liu, H. and **Pedatella, N.M.** (2021) 'Sudden stratospheric warming impacts on the ionosphere–thermosphere system: A review of recent progress', *Ionosphere dynamics and applications*, pp. 369–400. Available at: https://doi.org/10.1002/9781119815617.ch16

Hirschfeld, H.O. (1935) 'A connection between correlation and contingency', *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4), pp. 520–524. Available at: https://doi.org/10.1017/S0305004100013517

Hsu, C.Y., Soo, J.C., Lin, S.L., Wu, C.D., Chi, K.H., Hsu, W.C., Tseng, C.C. and **Chen, Y.C.** (2023) 'Using cluster algorithms with a machine learning technique and pmf models to quantify local-specific origins of PM2.5 and associated metals in Taiwan', *Environmental Pollution*, 316, p. 120652. Available at: https://doi.org/10.1016/j.envpol.2022.120652

Janches, D., Berezhnoy, A.A., Christou, A.A., Cremonese, G., Hirai, T., Horányi, M., Jasinski, J.M. and **Sarantos, M.** (2021) 'Meteoroids as one of the sources for exosphere formation on airless bodies in the inner solar system', *Space Science Reviews*, 217, 50, pp. 1–41. Available at: https://doi.org/10.1007/s11214-021-00827-6

Kogan, J. (2007) *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press.

Komorowski, M., Marshall, D.C., Salciccioli, J.D. and **Crutain, Y.** (2016) 'Exploratory data analysis', *Secondary Analysis of Electronic Health Records*, pp. 185–203. Available at: https://doi.org/10.1007/978-3-319-43742-2_15

Laštovička, J. (2023) 'Progress in investigating long-term trends in the mesosphere, thermosphere, and ionosphere', *Atmospheric Chemistry and Physics*, 23(10), pp. 5783–5800. Available at: https://doi.org/10.5194/acp-23-5783-2023

Li, G. and **Jung, J.J.** (2023) 'Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges', *Information Fusion*, 91, pp. 93–102. Available at: https://doi.org/10.1016/j.inffus.2022.10.008

Li, Z., Zhu, Y. and **Van Leeuwen, M.** (2023) 'A survey on explainable anomaly detection', *ACM Transactions on Knowledge Discovery from Data*, 18(1), pp. 1–54. Available at: https://doi.org/10.1145/3609333

Liang, R., Huang, C., Zhang, C., Li, B., Saydam, S. and **Canbulat, I.** (2023) 'The fusion of data visualisation and data analytics in the process of mining digitalisation', *IEEE Access* 11, pp. 40608–40628. Available at: https://doi.org/10.1109/ACCESS.2023.3267813

Lin, C., Labzovskii, L.D., Mak, H.W.L., Fung, J.C., Lau, A.K., Kenea, S.T., Bilal, M., Vande, H.J.D., Lu, X. and **Ma, J.** (2020) 'Observation of PM2.5 using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring', *Atmospheric Environment*, 227, p. 117410. Available at: https://doi.org/10.1016/j.atmosenv.2020.117410

Liu, X., Lu, D., Zhang, A., Liu, Q. and **Jiang, G.** (2022a) 'Data-driven machine learning in environmental pollution: gains and problems', *Environmental science & technology*, 56(4), pp. 2124–2133. Available at: https://doi.org/10.1021/acs.est.1c06157

Liu, Y., Tong, D., Cheng, J., Davis, S.J., Yu, S., Yarlagadda, B., Clarke, L.E., Brauer, M., Cohen, A.J., Kan, H. *et al.* (2022b) 'Role of climate goals and clean-air policies on reducing future air pollution deaths in China: a modelling study', *The Lancet Planetary Health*, 6(2), pp. e92–e99. Available at: https://www.osti.gov/servlets/purl/1855837

Lloyd, S.P. (1957) 'Least squares quantization in PCM', *Technical Report RR-5497, Bell Laboratories*. Available at: https://www.stat.cmu.edu/ rnugent/PCMI2016/papers/LloydKMeans.pdf

Lu, Q., Fu, H., Wang, R. and **Lu, S.** (2022) 'Collisionless magnetic reconnection in the magnetosphere', *Chinese Physics B*, 31(8), p. 089401. Available at: https://doi.org/10.1088/1674-1056/ac76ab

MacQueen, J.B. (1967) 'Some methods for classification and analysis of multivariate observations', *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1, pp. 281–297.

Mak, H.W.L. and **Lam, Y.F.** (2021) 'Comparative assessments and insights of data openness of 50 smart cities in air quality aspects', *Sustainable Cities and Society*, 69, p. 102868. Available at: https://doi.org/10.1016/j.scs.2021.102868

Mak, H.W.L., Laughner, J.L., Fung, J.C.H., Zhu, Q. and **Cohen, R.C.** (2018) 'Improved satellite retrieval of tropospheric NO_2 column density via updating of air mass factor (AMF): Case study of southern China', *Remote Sensing*, 10. Available at: https://doi.org/10.3390/rs10111789

Mak, H.W.L. and **Ng, D.C.Y.** (2021) 'Spatial and socio-classification of traffic pollutant emissions and associated mortality rates in high-density Hong Kong via improved data analytic approaches', *International Journal of Environmental Research and Public Health*, 18(12), p. 6532. Available at: https://doi.org/10.3390/ijerph18126532

Mumuni, A., Mumuni, F. and **Gerrar, N.K.** (2024) 'A survey of synthetic data augmentation methods in machine vision', *Machine Intelligence Research*, 21(5), pp. 831–869. Available at: https://doi.org/10.1007/s11633-022-1411-7

Mwitondi, K., Al Sadig, I., Hassona, R., Taylor, C. and **Yousef, A.** (2018a) 'Statistical estimate of radon concentration from passive and active detectors in Doha', *Data*, 3(3). Available at: https://doi.org/10.3390/data3030022

Mwitondi, K., Munyakazi, I. and **Gatsheni, B.** (2018b) 'Amenability of the United Nations Sustainable Development Goals to big data modelling', *International Workshop on Data Science-Present and Future of Open Data and Open Science, 12–15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall*, Mishima, Shizuoka, Japan.

Mwitondi, K., Munyakazi, I. and **Gatsheni, B.** (2020) 'A robust machine learning approach to SDG data segmentation', *Journal of Big Data*, 7(97). Available at: https://doi.org/10.1186/s40537-020-00373-y

Mwitondi, K.S., Moustafa, R.E. and **Hadi, A.S.** (2013) 'A data-driven method for selecting optimal models based on graphical visualisation of differences in sequentially fitted ROC model parameters', *Data Science Journal*, 12, pp. WDS247–WDS253. Available at: https://doi.org/10.2481/dsj.WDS-045

Mwitondi, K.S. and **Said, R.A.** (2013) 'A data-based method for harmonising heterogeneous data modelling techniques across data mining applications', *Journal of Statistics Applications & Probability*, 2(3), pp. 293–305. Available at: https://doi.org/10.12785/jsap/020312

Mwitondi, K.S. and **Said, R.A.** (2021) 'Dealing with Randomness and Concept Drift in Large Datasets', *Data*, 6(7). Available at: https://doi.org/10.3390/data6070077

Mwitondi, K.S. and **Zargari, S.A.** (2018) 'An iterative multiple sampling method for intrusion detection', *Information Security Journal: A Global Perspective*, 27(4), pp. 230–239. Available at: https://doi.org/10.1080/19393555.2018.1539790

Newby, D.E., Mannucci, P.M., Tell, G.S., Baccarelli, A.A., Brook, R.D., Donaldson, K., Forastiere, F., Franchini, M., Franco, O.H., Graham, I. *et al.* (2015) 'Expert position paper on air pollution and cardiovascular disease', *European Heart journal*, 36(2), pp. 83–93. Available at: https://doi.org/10.1093/eurheartj/ehu458

Omrani, N.E., Keenlyside, N., Matthes, K., Boljka, L., Zanchettin, D., Jungclaus, J.H. and **Lubis, S.W.** (2022) 'Coupled stratosphere-troposphere-atlantic multidecadal oscillation and its importance for near-future climate projection', *NPJ Climate and Atmospheric Science*, 5(1), p. 59. Available at: https://doi.org/10.1038/s41612-022-00275-1

Pan, Q., Harrou, F. and **Sun, Y.** (2023) 'A comparison of machine learning methods for ozone pollution prediction', *Journal of Big Data*, 10(1), p. 63. Available at: https://doi.org/10.1186/s40537-023-00748-x

Pika, A., ter Hofstede, A.H., Perrons, R.K., Grossmann, G., Stumptner, M. and **Cooley, J.** (2021) 'Using big data to improve safety performance: An application of process mining to enhance data visualisation', *Big Data Research*, 25, p. 100210. Available at: https://doi.org/10.1016/j.bdr.2021.100210

Ridzuan, F. and **Zainon, W.M.N.W.** (2022) 'Diagnostic analysis for outlier detection in big data analytics', *Procedia Computer Science*, 197, pp. 685–692. Available at: https://doi.org/10.1016/j.procs.2021.12.189

Shafiev, T. (2024) 'Development of a mathematical model and an efficient computational algorithm for predicting atmospheric pollution in industrial regions', in *AIP Conference Proceedings*, AIP Publishing. Available at: https://doi.org/10.1063/5.0199817

Sillmann, J., Aunan, K., Emberson, L., Büker, P., Van Oort, B., O'Neill, C., Otero, N., Pandey, D. and **Brisebois, A.** (2021) 'Combined impacts of climate and air pollution on human health and agricultural productivity', *Environmental Research Letters*, 16(9), p. 093004. Available at: https://doi.org/10.1088/1748-9326/ac1df8

Soares, P.H., Monteiro, J.P., Gaioto, F.J., Ogiboski, L. and **Andrade, C.M.G.** (2023) 'Use of association algorithms in air quality monitoring', *Atmosphere*, 14(4), p. 648. Available at: https://doi.org/10.3390/atmos14040648

Sun, Y. (2016) 'The changing role of China in global environmental governance', *Rising Powers Quarterly*, 1(1), pp. 43–53.

United Nations (2015) 'Sustainable Development Goals'. Available at: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

United Nations Environment Programme (UNEP) (2023) 'Annual report: Keeping the promise'. Available at: https://www.unep.org/annualreport/2023

Vardoulakis, S., Valiantis, M., Milner, J. and **ApSimon, H.** (2007) 'Operational air pollution modelling in the UK—street canyon applications and challenges', *Atmospheric Environment*, 41(22), pp. 4622–4637. Available at: https://doi.org/10.1016/j.atmosenv.2007.03.039

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.T., Aggarwal, C.C., Pei, J. and **Zhou, Y.** (2024) 'A comprehensive survey on data augmentation', *arXiv preprint arXiv:240509591*. Available at: https://doi.org/10.48550/arXiv.2405.09591

World Health Organization (2021) *Global Quality Guidelines: Particulate Matter (PM2.5 & PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, World Health Organisation, p. 290. Available at: https://www.who.int/publications/i/item/9789240034228

Wu, X., Wen, Q. and **Zhu, J.** (2024) 'Association rule mining with a special rule coding and dynamic genetic algorithm for air quality impact factors in Beijing, China', *PloS one*, 19(3), p. e0299865. Available at: https://doi.org/10.1371/journal.pone.0299865

Xu, M., Tian, W., Zhang, J., Screen, J.A., Zhang, C. and **Wang, Z.** (2023) 'Important role of stratosphere-troposphere coupling in the arctic mid-to-upper tropospheric warming in response to sea-ice loss', *npj Climate and Atmospheric Science*, 6(1), p. 9. Available at: https://doi.org/10.1038/s41612-023-00333-2

Xu, Y., Ho, H.C., Wong, M.S., Deng, C., Shi, Y., Chan, T.C. and **Knudby, A.** (2018) 'Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5', *Environmental pollution*, 242, pp. 1417–1426. Available at: https://doi.org/10.1016/j.envpol.2018.08.029

Yan, H., Cordier, M. and **Uehara, T.** (2024) 'Future projections of global plastic pollution: Scenario analyses and policy implications', *Sustainability*, 16(2), p. 643. Available at: https://doi.org/10.3390/su16020643

Zhang, B., Rong, Y., Yong, R., Qin, D., Li, M., Zou, G. and **Pan, J.** (2022a) 'Deep learning for air pollutant concentration prediction: A review', *Atmospheric Environment*, 290, p. 119347. Available at: https://doi.org/10.1016/j.atmosenv.2022.119347

Zhang, L. and **Yang, G.** (2022) 'Cluster analysis of PM2.5 pollution in China using the frequent itemset clustering approach', *Environmental Research*, 204, p. 112009. Available at: https://doi.org/10.1016/j.envres.2021.112009

Zhang, Q., Meng, X., Shi, S., Kan, L., Chen, R. and **Kan, H.** (2022b) 'Overview of particulate air pollution and human health in China: Evidence, challenges, and opportunities', *The Innovation*, 3(6). Available at: https://doi.org/10.1016/j.xinn.2022.100312

Zheng, S. and **Kahn, M.E.** (2017) 'A new era of pollution progress in urban China?', *Journal of Economic Perspectives*, 31(1), pp. 71–92. Available at: https://doi.org/10.1257/jep.31.1.71

Zhou, X., Hu, Y., Liang, W., Ma, J. and **Jin, Q.** (2021) 'Variational LSTM enhanced anomaly detection for industrial big data', *IEEE Transactions on Industrial Informatics*, 17(5), pp. 3469–3477. Available at: https://doi.org/10.1109/TII.2020.3022432

Zusman, E., Elder, M. and **Sussman, D.D.** (2020) *A Clean Air Sustainable Development Goal (SDG)*. Singapore: Springer Nature Singapore, pp. 1–12. Available at: https://doi.org/10.1007/978-981-15-2527-8_50-1

TO CITE THIS ARTICLE:
Mwitondi, K. and Mak, H.W.L. (2025) Robust Machine Learning Algorithmic Rules for Detecting Air Pollution in the Lower Parts of the Atmosphere. *Data Science Journal* 24: 27, pp. 1–24. DOI: https://doi.org/10.5334/dsj-2025-027

Submitted: 29 November 2024
Accepted: 28 August 2025
Published: 24 September 2025

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.